Bayesian Nonparametric Regression for Diabetes Deaths

Brian M. Hartman*
PhD Student, 2010
Texas A&M University
College Station, TX, USA

David B. Dahl
Assistant Professor
Texas A&M University
College Station, TX, USA

Abstract

Poisson regression models are commonly used in actuarial science to model count data (e.g., the number of claims) and covariates. The standard parametric formulation assumes that the response to covariates is constant across the population under study. Additionally, the shape of the posterior distribution is rather inflexible. In practice, however, the response to covariates may depend on a latent class and the posterior distribution may be highly nonparametric. We propose a Bayesian nonparametric model in which a Dirichlet process prior on the regression coefficients leads to a clustering of observations into groups having the same response to covariates. We illustrate the utility of our approach with an example dataset of diabetes deaths.

*Corresponding Author, Email: bhartman@stat.tamu.edu, Web: http://stat.tamu.edu/∼bhartman, Texas A&M University, MS 3143, College Station, TX 77843-3143

1 Introduction

As actuaries, we are constantly using information we know to estimate information we would like to know. For example, we could estimate the future claim experience of a policy by looking at the past claims, or various covariates. We have many tools to make those predictions, e.g. linear models, GLM's, time series analysis, splines, etc. But what if our data has some underlying structure? What if a simple linear model is the best fit, but the regression parameters are different for each of two subsets of the data? The method described in this paper is able to discover that underlying structure and improve both the understanding of the data and the future predictions.

1.1 Data Description

While the focus of this paper is an introduction to Dirichlet process priors (Ferguson, 1973), we use a simple dataset to illustrate the benefits of the method. The data in Table 1 contains the number of deaths due to diabetes in New South Wales, Australia in 2002, stratified by age and gender (De Jong and Heller, 2008).

Table 1: 2002 Diabetes Deaths in New South Wales

Gender	Age	Deaths	Population	Rate per 100K
Male	<25	3	1141100	0.26
Male	25-34	0	485571	0.00
Male	35-44	12	504312	2.38
Male	45-54	25	447315	5.59
Male	55-64	61	330902	18.43
Male	65-74	130	226403	57.42
Male	75-84	192	130527	147.10
Male	85+	102	29785	342.45
Female	<25	2	1086408	0.18
Female	25-34	1	489948	0.20
Female	35-44	3	504030	0.60
Female	45-54	11	445763	2.47
Female	55-64	30	323669	9.27
Female	65-74	63	241488	26.09
Female	75-84	174	179686	96.84
Female	85+	159	67203	236.60

2 Model

2.1 Sampling Model

We model the number of deaths (y_i) using a Poisson regression on the covariates x_i with exposure t_i:

$$\begin{aligned} y_i|\lambda_i &\sim Poi(y_i|\lambda_i) \\ \log\left(\lambda_i/t_i\right) &= x_i^T\beta \\ \lambda_i &= t_i \cdot \exp(x_i^T\beta). \end{aligned}$$

The link function allows us to rewrite the likelihood as:

$$\begin{aligned} y_i|\beta &\sim Poi(y_i|t_i \cdot \exp(x_i^T\beta)) \\ p(y_i|\beta) &= \frac{\exp\left\{-t_i \cdot \exp(x_i^T\beta)\right\}(t_i \cdot \exp(x_i^T\beta))^{y_i}}{y_i!}. \end{aligned}$$

Through independence, the joint probability is:

$$p(y|\beta) = \prod_{i=1}^{n} p(y_i|\beta).$$

2.2 Parametric Prior

When estimating this model in a Bayesian framework, we need to specify a prior distribution for β. A common choice is a multivariate normal distribution:

$$\beta \sim N_k(\beta_0, \Sigma_0).$$

With the prior and the likelihood specified, the posterior distribution follows from Bayes' rule:

$$\begin{aligned} p(\beta|y) &= \frac{p(\mathbf{y}|\beta)p(\beta)}{p(y)} \\ \log\left[p(\beta|y)\right] &\propto \sum_{i=1}^{n}\left[-t \cdot \exp(X_i^T\beta) + y_i\left(\log(t) + X_i^T\beta\right)\right] - \frac{1}{2}(\beta-\beta_0)^T\Sigma_0(\beta-\beta_0). \end{aligned}$$

The model can be fit using the Metropolis-Hastings algorithm (Hastings, 1970) or, better yet, the adaptive rejection sampler of Gilks et al. (1995) because the posterior is log-concave.

2.3 Nonparametric Prior

By using the above model, we are implicitly assuming that each observation has the same β vector. That may be naïve. As a simple example, let's assume that we only have the age covariate in our diabetes dataset. Age will likely have a different effect on males and females, but the standard parametric model does not allow that. We can enable our model to detect that substructure by simply adding an extra line to the prior specification

Original Parametric Model	Proposed Nonparametric Model
$y_i\|\beta \sim Poi(t_i \exp(x_i^T \beta))$	$y_i\|\beta_i \sim Poi(t_i \exp(x_i^T \beta_i))$
$\beta \sim G_0$	$\beta_i\|G \sim G$
	$G \sim DP(\alpha_0 G_0)$
$G_0 = N_k(\beta_0, \Sigma_0)$	$G_0 = N_k(\beta_0, \Sigma_0)$

where $DP(\alpha_0 G_0)$ is a Dirichlet process (Ferguson, 1973) with mass parameter α_0 and centering distribution G_0. This specification results in a Bayesian mixture model (Antoniak, 1974). For a review of these types of models, see Müller and Quintana (2004).

That simple addition allows each observation to have its own β vector while borrowing strength from the other observations through clustering.

3 Analysis

To illustrate, we fit the diabetes data with a Poisson regression model with link function

$$\lambda_i = t_i \cdot [\beta_0 + \beta_1 \text{age}]$$

where age is the midpoint of the age range for the observation. While this model is simple, it helps to solidify some abstract concepts. We will test if the method can model the gender effect without the gender information.

The parameters are estimated using a two step MCMC chain. The first step assigns each observation to a cluster and the second step returns a parameter vector for each cluster. We use the Auxiliary Gibbs scheme of Neal (2000).

For each iteration of the chain, we obtain a clustering and parameter estimates for each cluster. Each clustering could be different. For example, observations 12 and 17 may be in

the same cluster 80% of the time, 17 is with 18 10% of the time and 12 is never with 18. Presenting this information succinctly is an important task.

3.1 Posterior Clustering

A logical first step is to find a clustering of the observations which is optimal in some way. Binder (1978) was the first to purpose a loss function to minimize. More recent methods include least-square clustering from Dahl (2006) and max PEAR clustering from Fritsch and Ickstadt (2009). Applying the two newer methods to the diabetes data results in the same clustering presented in Table 2.

Table 2: 2002 Diabetes Deaths in New South Wales

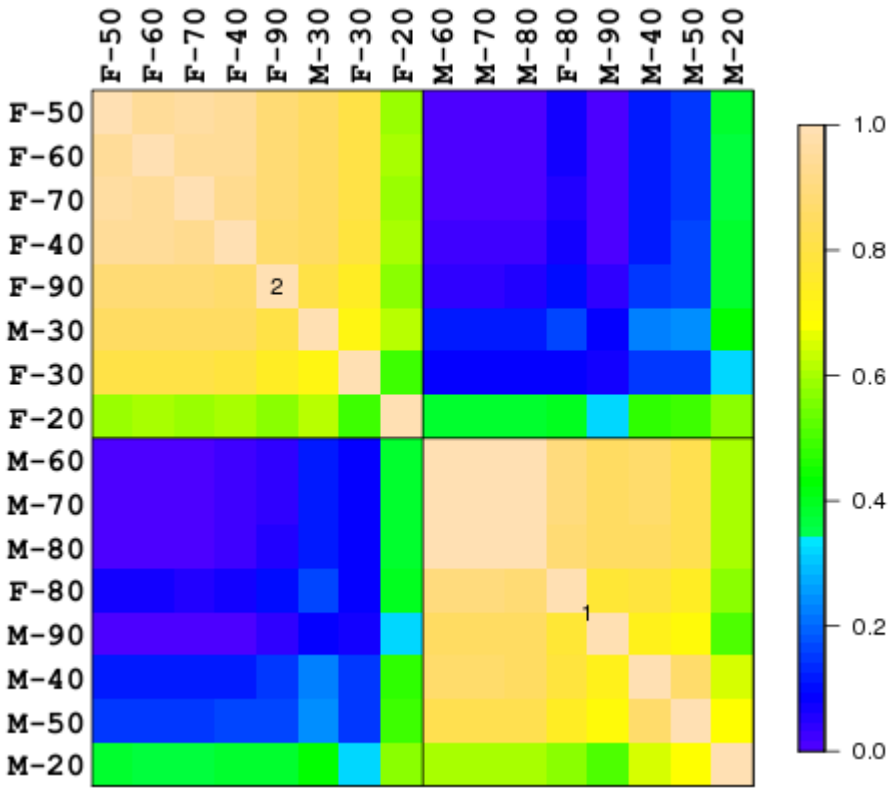
Gender	Age	Deaths	Population	Rate per 100K	Cluster
Male	<25	3	1141100	0.26	1
Male	25-34	0	485571	0.00	2
Male	35-44	12	504312	2.38	1
Male	45-54	25	447315	5.59	1
Male	55-64	61	330902	18.43	1
Male	65-74	130	226403	57.42	1
Male	75-84	192	130527	147.10	1
Male	85+	102	29785	342.45	1
Female	<25	2	1086408	0.18	2
Female	25-34	1	489948	0.20	2
Female	35-44	3	504030	0.60	2
Female	45-54	11	445763	2.47	2
Female	55-64	30	323669	9.27	2
Female	65-74	63	241488	26.09	2
Female	75-84	174	179686	96.84	1
Female	85+	159	67203	236.60	2

The clustering follows the gender difference in all but two cases, the 25-34 year-old males are clustered with the females and the 75-84 year-old females are clustered with the males. The death rate for males is higher at every age group except for the 25-34 year-olds. At that age, the death rates are very similar. We would expect for the males and the females to be clustered together. The reason for the 75-84 year-olds to be clustered together is not readily apparent.

3.2 Confidence Plots

While the posterior clustering gives a decent snapshot of the structure, there is uncertainty in the estimate. To examine that uncertainty, we examine the pairwise probability matrix. The element in the i^{th} row and j^{th} column is the proportion of draws where observations i and j were clustered together. Obviously, when $i = j$ the proportion equals one. A confidence plot (Dahl et al., 2009) is a heat map of the pairwise probability matrix. The rows and columns of the pairwise probability matrix are arranged to follow the posterior clustering. The color of each element represents the estimate of the pairwise probability. This graphic allows us to quickly assess which clusters are well defined and how the clusters are related to each other. Figure 1 is the confidence plot for our example dataset. F-40 stands for the female strata with an age midpoint of 40 (35-44 year-olds).

Figure 1: Confidence Plot



Notice that the two groups who were not classified with their gender (M-30 and F-80) are strongly classified in their cluster. The pairwise probabilities of those two observations are all greater than 0.5 within their cluster and less than 0.5 outside of it. There are two observations which have about a 0.5 probability of being clustered with any other observation, F-20 and M-20. The confidence plot gives us information about the strength of the clustering which

the posterior clustering lacks. Combining those two sources of information gives a strong picture of the underlying structure of the data.

4 Conclusion

Dirichlet process priors enable a modeler to more precisely fit a broad class of data structures. In our diabetes example, we were able to discover the gender effect without including the gender information. While it may be obvious in retrospect that gender would have an effect on the rate of diabetes deaths, there are other situations where the relationship may not be apparent a priori. Additionally, there may be a covariate that has an effect, but the data, or a good surrogate, is unavailable. For example, driving while talking on your cell phone increases the risk of an accident. Unfortunately, finding that data could prove difficult or impossible. If the effect is strong enough, using a Dirichlet process prior will allow the practitioner to find the structure and use it for prediction. Understanding the structure of the data would greatly improve ratemaking and pricing.

References

Antoniak, C. E. (1974). Mixtures of dirichlet processes with applications to bayesian nonparametric problems. *The Annals of Statistics 2*(6), 1152–1174.

Binder, D. (1978). Bayesian cluster analysis. *Biometrika 65*(1), 31.

Chib, S. and E. Greenberg (1995). Understanding the metropolis-hastings algorithm. *American Statistician 49*, 327–335.

Dahl, D. B. (2006). Model-based clustering for expression data via a dirichlet process mixture model. In M. V. Kim-Anh Do, Peter Müller (Ed.), *Bayesian Inference for Gene Expression and Proteomics*. Cambridge University Press.

Dahl, D. B., R. Day, and J. W. Tsai (2009). Distance-based probability distribution on set partitions with applications to protein structure prediction. Journal of the Royal Statistical Society: Series B, resubmitted.

De Jong, P. and G. Heller (2008). *Generalized linear models for insurance data*. Cambridge Univ Pr.

Ferguson, T. S. (1973). A bayesian analysis of some nonparametric problems. *The Annals of Statistics 1*(2), 209–230.

Fritsch, A. and K. Ickstadt (2009). Improved Criteria for Clustering Based on the Posterior Similarity Matrix. *Bayesian Analysis 4*(2), 367–392.

Gilks, W. and P. Wild (1992). Adaptive rejection sampling for Gibbs sampling. *Journal of the Royal Statistical Society. Series C (Applied Statistics) 41*(2), 337–348.

Gilks, W. R., N. G. Best, and K. K. C. Tan (1995). Adaptive rejection metropolis sampling within gibbs sampling. *Journal of the Royal Statistical Society. Series C (Applied Statistics) 44*(4).

Hastings, W. K. (1970). Monte carlo methods using markov chains and their applications. *Biometrika 57*, 97–109.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics 21*, 1087–1091.

Müller, P. and F. Quintana (2004). Nonparametric Bayesian data analysis. *Statistical science 19*(1), 95–110.

Neal, R. M. (2000, Jun). Markov chain sampling methods for dirichlet process mixture models. *Journal of Computational and Graphical Statistics 9*(2), 249–265.